DOCUMENT RESUME

ED 189 122 TM 800 287

AUTHOR Estes, Carole: Estes, Gary D.
TITLE Estimating a Correlation Coefficient Using a Multiple Matrix Sampling Disign.
PUB DATE Apr 80
NOTE 33p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Correlation: Elementary Secondary Education: Evaluation Methods: Grade 3: Grade 7: *Item Sampling: *Sampling: *Statistical Analysis: Statistical Bias: Test Reliability
IDENTIFIERS Comprehensive Test of Basic Skills: *Multiple Matrix Sampling: *Sample Size

ABSTRACT

Multiple matrix sampling is a sampling design in which both test items and examinees are randomly sampled from their respective populations. This study was designed to develop and assess a method for computing an estimate of a correlation coefficient when a multiple matrix sampling design is used. The examinee populations included 212 third-grade pupils and 232 seventh-grade pupils. The item populations consisted of the reading and mathematics subtests of the Comprehensive Tests of Basic Skills: 40, 45, 48, 25, and 25 items, respectively. The population parameters of interest were the correlation coefficients among the five subtests. Four matrix sample sizes were selected (2, 3, 4 and 5), and items and examinees were sampled. The correlation coefficients were adjusted using four different methods: (1) no adjustment: (2) validity-reliability, i.e. the matrix sample estimates and the subtest reliabilities were adjusted: (3) validity: and (4) reliability. The extent to which significant differences existed among the number of matrix samples, the adjustment method, and the interaction of these two factors were examined. The number of matrix samples affected the precision of the estimates. The adjustment method affected the estimate, the amount of bias in the estimate, and the precision of the estimate. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189122

ESTIMATING A CORRELATION COEFFICIENT USING A MULTIPLE MATRIX SAMPLING DESIGN

Carole Estes

Arizona State University

Gary D. Estes

Northwest Regional Educational Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Estes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, April 1980.

TM 800287

ESTIMATING A CORRELATION COEFFICIENT USING A MULTIPLE MATRIX SAMPLING DESIGN

Much of the literature on MMS can be organized by the way studies speak to the conjecture that MMS principles hold for all data. For example, some studies are of the type that show that the principles hold for a given set of data, some show that they hold for a lot of simulated data, and others attempt to prove that MMS principles always hold by using formal mathematical methods. Also included in the literature are applications of principles and expository pieces. A very brief review of multiple matrix sampling (MMS) terms is presented here so that the logic of the problem to be investigated will follow from the research that has accumulated to date.

Probably the clearest way to review MMS terminology is to depict certain relationships in the form of two-dimensional arrays or matrixes. In Figure 1, four different ways of approaching the task of obtaining information from a number of examinees on a test containing a specified number of items are shown.

Examinee sampling is the traditional alternative when it is impossible or not feasible to test every pupil with every item. Examinee sampling corresponds to a mixed model in analysis of variance terms because examinees constitute levels of the examinee factor which is sampled while the levels of the item factor are fixed. Another mixed model is item sampling in which levels of the item factor are sampled. This approach is not widely used, probably because information about items is more crucial than information about examinees in test

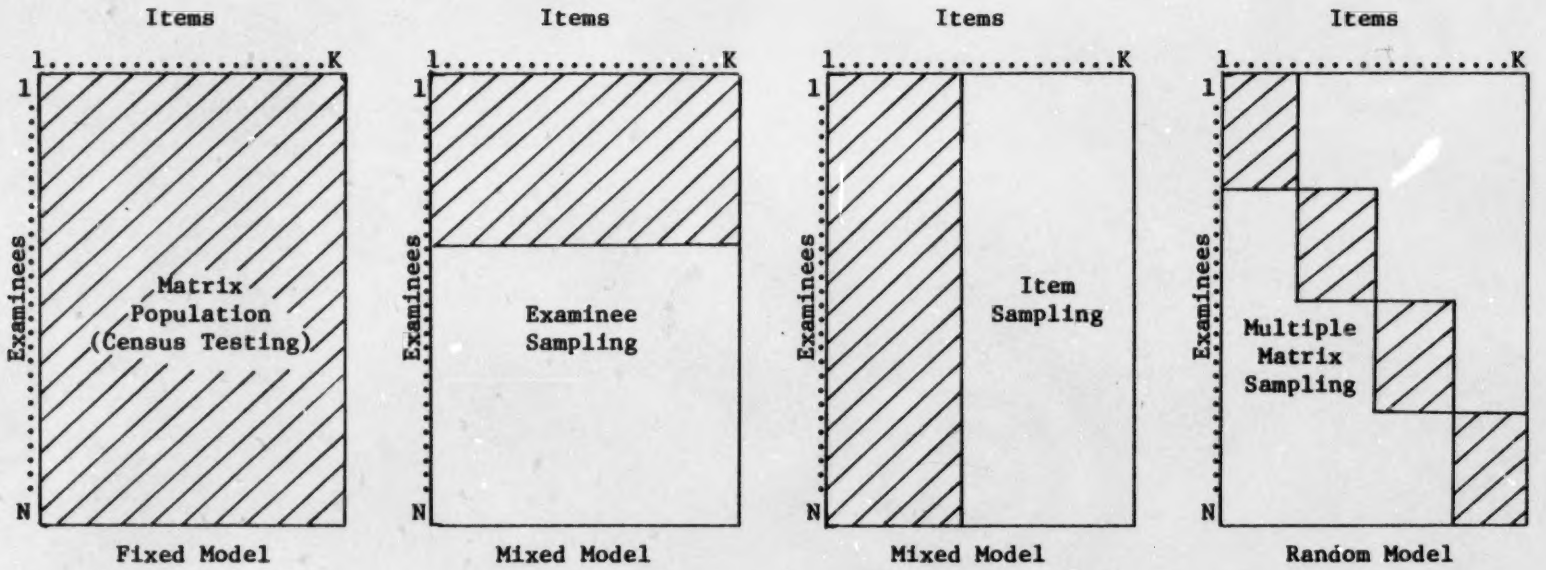


Figure 1. Four methods of sampling from an examinee x item matrix.

Table 1

Four t/k/n Sampling Fractions Replicated Over Two Grade Levels for the MMS Research Design

Grade Level	Number of Matrix Samples (t)	Number of Items Per Matrix Sample for Each Subtest (k)					Total Number of Items (K)[a]	Number of Examinees Per Matrix Sample (n)	Total Number of Examinees (N)[b]	Number of Random Replications
		Reading		Mathematics						
		Vocab.	Comp.	Comp.	Concepts	Applications				
3	2	20	22	24	12	12	180	106	212	50
	3	13	15	16	8	8	180	70	210	50
	4	10	11	12	6	6	180	53	212	50
	5	8	9	9	5	5	180	42	210	50
7	2	20	22	24	12	12	180	116	232	50
	3	13	15	16	8	8	180	77	231	50
	4	10	11	12	6	6	180	58	232	50
	5	8	9	9	5	5	180	46	230	50

[a]Since K was actually 183 (40 + 45 + 48 + 25 + 25) and therefore not an integral multiple of k, items were randomly deleted to produce equal-sized matrix samples of size k for each of the 50 random replications.

[b]N was actually 212 and 232 for the two grade levels and was handled in the same manner as K.

construction. MMS is called a random model because both items and examinees are randomly sampled from their respective populations. This particular MMS design is also nonoverlapping. That is, both items and examinees are sampled without replacement. It is exhaustive because all items and examinees are eventually sampled. The most common applications of MMS are of this type and illustrate the general principle that if one wants to estimate the mean of the total group on all the test items, the MMS mean estimates tend to be closer to the actual population mean and closer to each other than are those of the examinee sampling process.

Lord (1962) contributed one of the first studies in MMS in which he estimated the mean, variance and distribution of scores with a 10/7/100 sampling plan based on a test of 70 items and total group of 1,000 examinees.[1] Plumlee (1964) and Cook and Stufflebeam (1967) supported Lord's findings using without-replacement designs. Pugh (1971) varied sample size and concluded with the previous authors that MMS procedures produced good estimates of the mean and adequate estimates of the variance. MMS procedures and examinee sampling procedures gave equally effective estimates for Owens and Stufflebeam (1969). Sirotnik, in 1970 and 1974, showed that scrambling test content among several different tests has minimal or no effect on mean examinee performance, variance of scores, variance of item difficulties, and K-R 20. Further research suggested that context effects were minimal (French & Greer, 1964; Huck & Bowers, 1972; Marso, 1970; Sax and Cromack, 1966).

Knapp (1968) provided a pioneer effort in MMS designs by estimating the mean, variance, and K-R 20 with a with-replacement or BIB design.

---

[1]The notation 10/7/100 is used when an MMS sampling plan is described. The general term is t/k/n, where t = the number of matrix samples, k = the number of items per matrix sample, and n = the number of examinees per matrix sample.

Barcikowski (1972) conducted a study using the Monte Carlo approach in looking at the effects of differing difficulty and discrimination indexes on two types of sampling. He found that item-examinee sampling (MMS) gave better estimates of the mean for all combinations of difficulty and discrimination indexes but examinee sampling gave better estimates of test variance for certain difficulty-discrimintion combinations. In a later study (1974), he found that an MMS plan that is optimal for estimating the total test mean may be inefficient for estimating the variance and vice versa.

Shoemaker (1970, 1971, 1972, 1973) has provided most of the computer simulation studies. His conclusions include: that the total number of observations (item-examinee data points) appear to be the critical factor in the efficiency of MMS designs; that if this total number of observations were fairly large, e.g., 1,500, the number of matrix samples, item sample size, and examinee sample size made very little difference; and that a sampling plan that minimizes the standard error of estimate for the mean will also minimize generally the standard error of estimate for the variance, and vice versa.

The problem that is undertaken in this study is the development and assessment of a method for computing an estimate of a correlation coefficient when a multiple matrix sampling (MMS) design is used. MMS is a relatively recent technique that has shown promise for providing solutions to the problem of feasibility and cost of testing every pupil with every test item when evaluating school programs.

The questions that will be answered by this study are:

1. Can a procedure for estimating a correlation coefficient be successfully applied when a without-replacement[2] MMS design is taken

---

[2]Examinees and items occurring in one matrix sample do not occur in another.

that is equal-sized[3] for both examinees and items and exhaustive for both populations?

2. If an MMS design can be applied and a procedure can be used to estimate a correlation, what differences arise when the MMS plans are varied together with four computational methods, only one of which makes use of reliabilities computed from the shortened tests?

3. Does examination of the frequency distributions obtained by plotting replications by an MMS-adjustment method yield information about the bias and precision of estimates based on different MMS plans and adjustment method combinations?

4. Can the MMS plan estimates be replicated when the examinee population is varied?

METHOD

In this MMS design, the examinee populations equaled 212 third grade pupils and 232 seventh grade pupils, i.e., N = 212 and 232. The item populations equaled 40, 45, 48, 25 and 25, i.e., K = 40, 45, 48, 25 and 25, respectively, the Reading and Math subtests of the Comprehensive Tests of Basic Skills. The population parameters of interest were the correlation coefficients among the five subtests. Four matrix sample designs were applied to the population matrix in the following manner:

1. Four matrix sample sizes were selected, i.e., t = 2, 3, 4 and 5, so that a range of numbers of matrixes could be examined comprising a sufficient number of items and examinees per matrix sample. Table 1 presents the numbers of items and examinees included in each t/k/n/ plan.

---

[3] Equal-sized matrix samples are those in which the same number of examinees or items occur in each matrix sample.

2. A computer program was written which assigned a random number to each examinee in the grade level to be analyzed and selected the required number of examinees for each matrix sample using a random number table.

3. The computer program also assigned a random number to each item in a subtest, selected the required numbers of items for each matrix sample using a random number table and summed the resulting dichotomously scored items to arrive at the previously selected examinee's shortened subtest score.

4. All examinees were reassigned new random numbers and the required number of examinees for each matrix sample was selected using a random number table. All items in a subtest were reassigned new random numbers, the required number of items for each matrix sample was selected using a random number table, and the resulting items summed to arrive at each previously selected examinee's subtest score.

5. The above procedure was repeated so that examinees and items were each randomly selected for inclusion in matrix samples a total of 50 times for each matrix sample size, i.e., two, three, four, and five. The procedure was repeated 50 times because it was thought that this number of random replications would probably produce a fairly stable mean estimate.

6. The following descriptive statistics were computed for each subtest in the population matrix and for each subtest in each of the matrix samples: (a) the mean of the subtest scores, (b) the standard deviation of the subtest scores, (c) the mean of each item score in the

subtest, (d) the variance of each item score in the subtest, and (e) the intercorrelations among all subtests.

7. Kuder-Richardson 20 reliability estimates were then computed for each subtest.

8. The correlation coefficients computed among the subtests were adjusted using four different methods: (a) No addjustment: the matrix sample estimates were summed and divided by the number of matrix samples; (b) Validity-reliability: the matrix sample estimates were adjusted because they were based on only a portion of the items of the original subtest by the following formula given by Spearman:

$$R_{PQ} = \frac{r_{11}}{\sqrt{\frac{1}{P} + \left(1 - \frac{1}{P}\right) r_{11}} \sqrt{\frac{1}{Q} + \left(1 - \frac{1}{Q}\right) r_{II}}}$$

P = the number of times one subtest is increased in length

Q = the number of times the other subtest is increased in length

r_{11} = the reliability of one subtest

r_{II} = the reliability of the other subtest

In addition, the subtest reliabilities were also adjusted for the fact that they were based on a shortened test by the general Spearman-Brown formula:

$$R_{PP} = \frac{Pr_{11}}{1 + (P - 1)\, r_{11}}$$

(c) Validity: the matrix sample estimates were adjusted as in (b) except that the K-R 20's were not adjusted. This method is the one proposed in most textbooks (Gulliksen, 1950; Lord and Novick, 1968).

(d) Reliability: the matrix sample estimates were adjusted using the Spearman-Brown reliability formula in (b) and replacing r_{11} with the correlation coefficient between subtests.

9. The adjusted correlation coefficients were then converted to Fisher z coefficients using the following formula (Hays & Winkler, 1979):

$$z = 1/2 \log \left[(1 + r)/(1 - r)\right]$$

10. The Fisher z coefficients of multiple matrix sample r's within each t/k/n-adjustment method plan were averaged. From this mean was subtracted the z coefficient of the population r. The result of this subtraction was the difference between the MMS estimate and the population value, i.e., the estimate-population difference.

11. The cumulative frequency distributions of the estimate-population differences resulting from the 50 random replications of a particular matrix sample size, adjustment method, and subtest pairing were plotted and the mean and variance of the frequency distribution were calculated.

12. The means and variances of the distribution of the z value of these discrepancies based on each t/k/n plan and adjustment method were comipared to the population correlation coefficient $\pm$ 1.96 standard deviations.

13. The extent to which significant differences existed among Factor A, the number of matrix samples; Factor B, the adjustment method; and Factor A x B, the interaction of factors, was examined by calculating a two-way ANOVA with replications by subtest pair and grade level. That is, differences among estimates produced by two, three, four and five matrix sample sizes were examined. Also, differences among estimates produced by each of four adjustment methods were examined. Finally, differences among estimates produced by each of 16 combinations of sample size and adjustment method were examined.

14. The extent to which bias existed among the estimate-population differences was examined by calculating the 95% confidence interval about the population values and comparing the mean estimate-population differences to this interval.

15. The extent of the precision among the estimates from the four matrix sample sizes was examined by calculating Hartley's F_{max} statistic for determining homogeneity of variances.

RESULTS

The results of the data analyses are organized around the following three topics of investigation:

1. The differences among estimate-population difference means were examined.

2. The estimate-population difference means were examined to see how many are equal to or less than the 95% confidence interval, i.e., an analysis of the bias among the estimate means was made.

3. The differences among estimate-population difference variances were examined, i.e., an analysis of the precision among the estimate means was made.

The efficiency of the multiple matrix sampling estimates of ten correlation coefficients using four adjustment methods are summarized below.

1. Differences among estimates: (a) Different numbers of matrix samples yield estimates of different accuracies, disregarding adjustment method. However, less than 1% of the variance was accounted for. (b) Estimates differ significantly according to the method of adjustment,

disregarding number of matrix samples. The order of the means from largest underestimate to largest overestimate is: no adjustment, validity-reliability, validity, and reliability. Seventy to 80% of the variance is accounted for. (c) The accuracy of estimates varies according to the particular combination of adjustment method, but with the same order of means as (1b). Seven to 12 percent of the variance is accounted for by the interaction. (See Tables 2-11 in the Appendix.)

2. Analysis of bias: (a) Estimates are unbiased for the different numbers of matrix samples, disregarding adjustment method. (b) Estimates are usually unbiased for the "validity adjustment method" and often unbiased for the "validity-reliability adjustment method," disregarding number of matrix samples. (c) Estimates are usually unbiased for different "number of matrix samples" and "adjustment method" combinations: validity adjustment for two and three samples, validity-reliability adjustment for two and three samples, and no adjustment for two samples. Reliability adjustments were usually biased and led to overestimates.

3. Analysis of precision: (a) At $t = 2$, there were significant differences in precision of the estimates in .05 of cases; at $t = 3$, in .75 of cases; at $t = 4$, in .95 of cases; and at $t = 5$, in 1.00 of cases. (b) The ordering of the variances is that the "no adjustment method" resulted in the greatest precision and the "validity method" resulted in the least precision and yielded precisions which decreased the most rapidly from $t = 2$ through $t = 5$.

4. Replicability analysis: The above trends were consistent across grade levels.

INTERPRETATIONS AND CONCLUSIONS

Interpretations of the results of this study will be organized around explanations offered for these two questions:

1. What properties of the three adjustment methods are responsible for the significant differences in their estimates, degree of bias, and degrees of precision?

2. What are the decisions a researcher must make the consequences of these decisions when choosing a sampling plan and corresponding adjustment method for estimating a correlation coefficient?

Properties of the Adjustment Methods

Because each adjustment method differed in the way test length and test reliability were used in their formulas, they produced different estimates of the correlation coefficients. Additionally, the relatively large percent of criterion variance accounted for by the method of adjustment together with the interaction factor was evidence of their strength and importance in this experimental desingn.

No adjustment method. This method consistently produced underestimates of the correlation coefficient because it did not take into account and correct for the decrease in reliability of a shortened test. Further support for this explanation was seen in the increasing bias of the underestimates when the number of matrix samples increased from 3 to 4 to 5.

Validity-reliability and validity adjustment methods. These two methods are discussed together because the methods were identical except for the reliability adjustment in the validity-reliability method. Both employed the number of times a test was increased in length and the

reliabilities of the tests. However, where the validity method used the observed K-R 20's of the shortened test directly, the validity-reliability method first adjusted the observed K-R 20's with the Spearman-Brown formula, due to the fact that they were based on shortened tests. The results of the "validity-reliability adjustment method" tended to result in underestimates and the "validity adjustment method" tended to result in overestimates. Guilford (1965) noted:

> If either r_{yy} or r_{xx} is underestimated, the corrected r_{yx} will be overestimated. If either reliability coefficient is overestimated, the corrected r_{yx} will be underestimated. It is probably best, if one wishes to be on the conservative side, that if anything, a reliability estimate should be too large when used for this purpose. On the other hand, it is likely that most estimates of internal-consistency reliability are too low, which is in the wrong direction for conservatism. (p. 488)

Lord and Novick (1968) also mention this characteristic of the validity adjustment:

> The quantities ρ_{xx1} and ρ_{yy1} appear in the denominator.... Hence, if these values are understated, the value of the ratio will be overstated. In practice, this is what all too often occurs when the correlation between x and y is estimated on the basis of a single administration of each of the measurements x and y in close temporal contiguity. In this case, the error variation present in the estimate is the error due to the imprecision of each of the two measurements....If we use an internal-analysis method, then a lack of homogeneity of the items (or subtests) will...result in an underestimation of reliability. Hence the correction for attenuation tends to be an overcorrection of unknown extent....Unfortunately this procedure is often unsatisfactory. Indeed in some cases this approach has led to the impossible conclusion that $\rho(T_x, T_y) > 1$. (p. 138)

This last statement was found to be true for the validity method in this study and accounts for the fact that the variances of estimate-population differences become larger at a much faster rate than the other methods as t approaches 5. As t approaches 5, the distribution

spreads out as compared to the other methods. Thus, the "validity adjustment method" may have provided the most unbiased estimates for two through five matrix samples compared to the other methods, but its precision decreased quickly after two matrix samples. The "validity-reliability adjustment method" provided unbiased estimates for two matrix samples and usually for three matrix samples, but its precision decreased relatively slowly after two matrix samples.

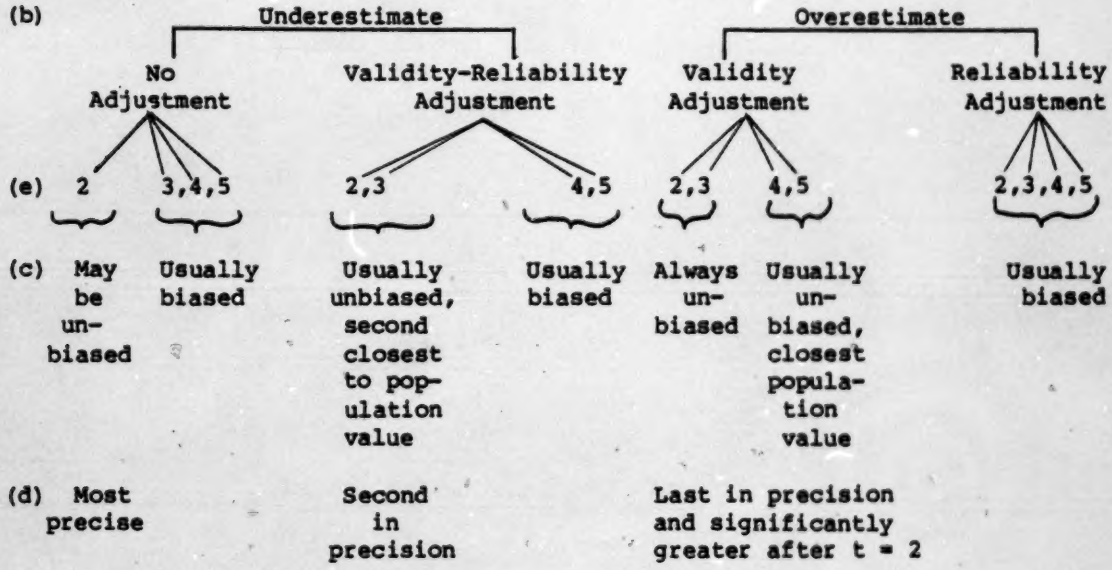
Reliability adjustment method. This method consistently produced overestimates of the correlation coefficient because it was based on the Spearman-Brown formula that provides the reliability of a lengthened test. In comparing the procedures for the effect of increased test length on reliability-validity, Lord and Novick (1968) noted "that validity increases much more slowly with length than reliability" (p. 115). Thus, the validity estimate was adjusted as though it were an internal consistency reliability estimate and the resultant adjustment followed the general tendency to increase as noted by Lord and Novick.

Decisions and Their Consequences in Choosing Estimate Plans

A researcher who wants to estimate a population correlation coefficient probably should consider some of the following decisions: (a) What is the purpose of estimating the correlation coefficient? (b) Would an underestimate or an overestimate be more acceptable for the particular study? (c) How much of an under- or overestimte (degree of bias) would be acceptable? (e) How many matrix samples are desired?

In examining the purpose for estimating a correlation coefficient, an education evaluator needs to determine the decisions which will be based on the results of the strength of relation found. For example,

validation procedures for a criterion-referenced test may include its relation to a norm-referenced test and one of the uses of the former may be to assess pupil mastery of skills and placement in various curricula. In contrast, an evaluator's purpose for estimating may be limited to including the relation between two variables as a secondary part of a larger study and not the major point of interest. This would be the case where the descriptive part of an experimental study would involve the relation among categorical variables, e.g., socioeconomic status, IQ. Based on the findings of this study, the outcomes of decisions (b) through (e) are presented below.



In general, the most reasonable decision would seem to be that of imposing the validity adjustment with two matrix samples. However, if the amount of testing time must be decreased, the validity-reliability adjustment with three matrix samples is an acceptable alternative.

The following conclusions might be drawn from the current study. However, the reader should be aware of the fact that no hard and fast

rules have surfaced from this one investigation. It will have to be replicated by others and alternate methods attempted for estimating correlation coefficients before any definite guidelines can be presented.

1. An estimate of test reliability and the number of times a test is lengthened are two variables which should probably be included in a method for estimating a correlation coefficient by means of multiple matrix sampling.

2. The number of matrix samples used interacts with the method of estimation used in such a way as to produce unbiased to increasingly imprecise estimates as the number of matrix samples increases.

3. Although the validity method may provide a more accurate estimate of the population value, as compared to the validity-reliability method, it is not as precise an estimate.

Before practitioners select a method, they will need to make decisions about the purposes of estimating the correlation coefficient and about the degrees and directions of bias and precision which are tolerable. Finally, it needs to be said once again that practitioners should recognize that the above conclusions do not define specific guidelines for estimating correlation coefficients, but rather are an initial attempt at shedding some light on the problem of developing an empirical method for estimating relationships among variables by means of multiple matrix sampling.

REFERENCES

Barcikowski, R. S. A Monte Carlo study of item sampling (versus traditional sampling) for norm construction. Journal of Educational Measurement, 1972, 9, 209-214.

Barcikowski, R. S. The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. Educational and Psychological Measurement, 1974, 34, 231-237.

Cook, D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Journal of Educational Measurement, 1967, 4, 27-33.

French, J. L., & Greer, D. Effect of test-item arrangement on physiological and psychological behavior in primary-school children. Journal of Educational Measurement, 1964, 1, 151-153.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1965.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hays, W. L., & Winkler, P. L. Statistics: Probability, inference, and decision. New York: Holt, Rinehart, & Winston, 1971.

Huck, S. W., & Bowers, N. D. Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 1972, 9, 105-111.

Knapp, T. R. An application of balanced incomplete block designs to the estimation of test norms. Educational and Psychological Measurement, 1968, 28, 265-272.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Marso, R. Test item arrangement, testing time, and performance. Journal of Educational Measurement, 1970, 7, 113-118.

Owens, T. R., & Stufflebeam, D. L. An experimental comparison of item sampling and examinee sampling for estimating test norms. Journal of Educational Measurement, 1969, 6, 75-83.

Plumlee, L. B. Estimating means and standard deviations from partial data--an empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Pugh, R. C. Empirical evidence on the application of Lord's sampling technique to Likert items. Journal of Experimental Education, 1971, 39, 54-56.

Sax, G., & Cromack, T. R. The effects of various forms of item arrangements on test performance. Journal of Educational Measurement, 1966, 3, 309-311.

Shoemaker, D. M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970, 7, 255-262.

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.

Shoemaker, D. M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. Educational and Psychological Measurement, 1972, 32, 705-714.

Shoemaker, D. M. A note on allocating items to subtests in multiple matrix sampling and approximating standard errors of estimate with the jackknife. Journal of Educational Measurement, 1973, 10, 211-219.

Sirotnik, K. An investigation of the context effect in matrix sampling. Journal of Educational Measurement. 1970, 7, 199-208.

Sirotnik, K., & Wellington, R. Scrambling content in achievement testing: An application of multiple matrix sampling in experimental design. Journal of Educational Measurement, 1974, 11, 179-188.

APPENDIX

Table 2

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences Computed from N = 50 Random Replications Compared to Population Values: Vocabulary-Comprehension Correlation Coefficient

Adjustment Method	Number of Matrix Samples							
	2		3		4		5	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Third Grade								
no	-.089	.034	-.185	.045*	-.281	.054*	-.332	.054*
validity-reliability	-.013	.040	-.097	.052	-.200	.064	-.252	.065
validity	.068	.047	.113	.081	.133	.151	.323	.289
reliability	.228	.037	.314	.050	.340	.063	.384	.065
	$\underline{z}$ = 1.163, $C.I._{.95}$ = .136, r = .82							
Seventh Grade								
no	-.148	.047	-.262	.047*	-.359	.055*	-.407	.058*
validity-reliability	-.064	.055	-.168	.059	-.273	.067	-.323	.071
validity	.025	.066	.053	.103	.077	.189	.264	.333
reliability	.161	.055	.222	.054	.241	.067	.283	.073
	$\underline{z}$ = 1.122, $C.I._{.95}$ = .130, r = .81							

*$\underline{p} < .05$

Table 3

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences Computed from N = 50 Random Replications Compared to Population Values: Vocabulary-Computation Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
				Third Grade				
no	-.102	.032	-.163	.041	-.205	.044*	-.258	.048*
validity-reliability	-.064	.036	-.115	.046	-.155	.050	-.208	.055
validity	-.026	.039	-.014	.060	.014	.080	.050	.151
reliability	.190	.036	.301	.048	.377	.054	.410	.061
			$\underline{z}$ = .904, C.I.$_{.95}$ = .136, r = .72					
				Seventh Grade				
no	-.064	.029	-.126	.038	-.177	.043*	-.213	.047*
validity-reliability	-.028	.031	-.082	.042	-.132	.048	-.167	.052
validity	.007	.034	.006	.052	.008	.070	.033	.087
reliability	.217	.033	.318	.048	.380	.057	.425	.067
			$\underline{z}$ = .771, C.I.$_{.95}$ = .130, r = .65					

*$\underline{p}$ < .05

Table 4

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Vocabulary-Concepts Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Third Grade								
no	-.173	.037	-.255	.053*	-.311	.047*	-.348	.052*
validity-reliability	-.111	.042	-.181	.061	-.234	.057	-.260	.107
validity	-.051	.048	-.028	.085	.047	.169	.187	.267
reliability	.113	.041	.195	.064	.250	.061	.298	.071
	$\underline{z}$ = .922, $C.I._{.95}$ = .136, r = .73							
Seventh Grade								
no	-.133	.031	-.221	.048*	-.272	.050*	-.325	.062*
validity-reliability	-.072	.034	-.148	.058	-.198	.062	-.253	.072
validity	-.014	.039	.004	.084	.058	.145	.114	.220
reliability	.157	.035	.224	.059	.286	.065	.310	.089
	$\underline{z}$ = .869, $C.I._{.95}$ = .130, r = .70							

*$\underline{p} < .05$

Table 5

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Vocabulary-Application Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
				Third Grade				
no	-.082	.038	-.164	.038*	-.234	.040*	-.274	.045*
validity-reliability	-.024	.042	-.096	.044	-.166	.046	-.207	.054
validity	.033	.048	.052	.062	.074	.087	.200	.220
reliability	.216	.041	.306	.044	.351	.050	.401	.058
				$\underline{z}$ = .935, C.I.$_{.95}$ = .136, r = .73				
				Seventh Grade				
no	-.093	.036	-.168	.049*	-.230	.044*	-.263	.053*
validity-reliability	-.044	.040	-.108	.056	-.170	.049	-.206	.065
validity	.004	.045	.015	.078	.025	.077	.084	.208
reliability	.190	.040	.278	.060	.328	.057	.378	.073
				$\underline{z}$ = .822, C.I.$_{.95}$ = .130, r = .68				

*$\underline{p}$ < .05

Table 6

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Comprehension-Computation Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Third Grade								
no	-.066	.032	-.103	.037	-.135	.043	-.176	.053*
validity-reliability	-.040	.034	-.070	.040	-.099	.047	-.139	.057
validity	-.015	.036	-.002	.046	.015	.062	.026	.088
reliability	.216	.036	.350	.044	.437	.055	.479	.072
	$\underline{z}$ = .787, $C.I._{.95}$ = .136, r = .66							
Seventh Grade								
no	-.089	.028	-.152	.038	-.209	.055*	-.239	.053*
validity-reliability	-.043	.031	-.098	.042	-.154	.062	-.181	.061
validity	.001	.034	.013	.053	.029	.095	.116	.161
reliability	.200	.031	.305	.045	.361	.069	.421	.071
	$\underline{z}$ = .850, $C.I._{.95}$ = .130, r = .69							

*$\underline{p} < .05$

Table 7

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Comprehension-Concepts Correlation Coefficient

Adjustment Method	Number of Matrix Samples							
	2 Mean	2 Standard Deviation	3 Mean	3 Standard Deviation	4 Mean	4 Standard Deviation	5 Mean	5 Standard Deviation
Third Grade								
no	-.180	.041	-.249	.043*	-.303	.049*	-.340	.056*
validity-reliability	-.125	.046	-.184	.050	-.234	.058	-.260	.101
validity	-.073	.052	-.054	.071	-.000	.123	.104	.254
reliability	.105	.046	.201	.053	.261	.063	.307	.074
	$\underline{z}$ = .918, C.I.$_{.95}$ = .136, r = .72							
Seventh Grade								
no	-.155	.033	-.240	.050*	-.319	.055*	-.356	.069*
validity-reliability	-.088	.039	-.161	.061	-.242	.068	-.279	.080
validity	-.023	.046	.006	.092	.015	.144	.133	.279
reliability	.131	.037	.207	.061	.234	.074	.278	.099
	$\underline{z}$ = .899, C.I.$_{.95}$ = .130, r = .72							

*$\underline{p} < .05$

Table 9

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Computation-Concepts Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
				Third Grade				
no	-.122	.041	-.177	.047	-.235	.047*	-.293	.063*
validity-reliability	-.068	.046	-.112	.052	-.165	.057	-.216	.118
validity	-.018	.052	.021	.067	.074	.141	.127	.265
reliability	.166	.046	.278	.058	.333	.061	.354	.088
			$\underline{z}$ = .877, C.I.$_{.95}$ = .136, r = .70					
				Seventh Grade				
no	-.131	.041	-.236	.045*	-.320	.058*	-.377	.058*
validity-reliability	-.054	.047	-.148	.056	-.235	.073	-.292	.071
validity	.022	.055	.044	.088	.073	.171	.181	.242
reliability	.170	.045	.235	.053	.266	.072	.293	.075
			$\underline{z}$ = 1.014, C.I.$_{.95}$ = .130, r = .77					

*$\underline{p} < .05$

Table 10

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Computation-Applications Correlation Coefficient

Adjustment Method	Number of Matrix Samples							
	2		3		4		5	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Third Grade								
no	-.077	.038	-.151	.050*	-.210	.044*	-.260	.047*
validity-reliability	-.027	.042	-.090	.056	-.149	.049	-.100	.054
validity	.024	.048	.040	.076	.058	.080	.110	.133
reliability	.225	.041	.326	.058	.388	.053	.429	.062
	$\underline{z}$ = .965, C.I.$_{.95}$ = .136, r = .75							
Seventh Grade								
no	-.111	.044	-.203	.045*	-.266	.048*	-.314	.054*
validity-reliability	-.050	.049	-.132	.053	-.191	.056	-.245	.063
validity	.012	.055	.021	.074	.080	.113	.146	.194
reliability	.191	.048	.272	.053	.329	.060	.370	.070
	$\underline{z}$ = 1.001, C.I.$_{.95}$ = .130, r = .76							

*$\underline{p} < .05$

Table 11

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences

Computed from N = 50 Random Replications Compared to Population Values:

Concepts-Applications Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
				Third Grade				
o	-.130	.039	-.222	.058*	-.276	.046*	-.316	.050*
alidity-reliability	-.064	.045	-.148	.069	-.196	.056	-.228	.104
alidity	.001	.052	.007	.100	.103	.186	.220	.297
eliability	.157	.043	.226	.070	.286	.056	.329	.068
				$\underline{z}$ = .886, C.I.$_{.95}$ = .136, r = .71				
				Seventh Grade				
o	-.185	.045*	-.330	.040*	-.410	.062*	-.471	.057*
alidity-reliability	-.073	.057	-.212	.051	-.293	.082	-.364	.075
alidity	.048	.075	.090	.156	.323	.362	.427	.437
eliability	.122	.057	.146	.046	.182	.075	.208	.073
				$\underline{z}$ = 1.140, C.I.$_{.95}$ = .130, r = .81				

*$\underline{p}$ < .05

Table 8

A x B Cell Means and Standard Deviations of Estimate Minus Population Differences Computed from N = 50 Random Replications Compared to Population Values: Comprehension-Applications Correlation Coefficient

	Number of Matrix Samples							
	2		3		4		5	
Adjustment Method	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Third Grade								
no	-.077	.037	-.137	.036*	-.204	.046*	-.230	.058*
validity-reliability	-.027	.041	-.077	.042	-.144	.052	-.167	.069
validity	.022	.046	.051	.059	.064	.090	.170	.204
reliability	.221	.041	.336	.042	.386	.056	.456	.073
	$\underline{z}$ = .929, C.I.$_{.95}$ = .136, r = .73							
Seventh Grade								
no	-.111	.034	-.191	.046*	-.249	.059*	-.289	.054*
validity-reliability	-.058	.038	-.130	.053	-.186	.068	-.229	.063
validity	-.007	.042	-.004	.073	.026	.114	.073	.190
reliability	.172	.039	.253	.057	.303	.078	.346	.076
	$\underline{z}$ = .833, C.I.$_{.95}$ = .130, r = .68							

*$\underline{p} < .05$